Spanish Sentiment Analysis using Universal Language Model Fine-Tuning: a Detailed Case of Study*

Daniel Palomino[1 [0000−0003−2075−3379]] and
José Ochoa-Luna[1 [0000−0002−8979−3785]]

Department of Computer Science
Universidad Católica San Pablo, Arequipa, Perú

Abstract. Transfer Learning has emerged as one of the main image classification techniques for reusing architectures and weights trained on big datasets so as to improve small and specific classification tasks. In Natural Language Processing, a similar effect is obtained by reusing and transferring a language model. In particular, the Universal Language Fine-Tuning (ULMFiT) algorithm has proven to have an impressive performance on several English text classification tasks. In this paper, we aim at improving current state-of-the-art algorithms for Spanish Sentiment Analysis of short texts. In order to do so, we have adapted a ULMFiT algorithm to this setting. Experimental results on benchmark datasets show the potential of our approach.

Keywords: Sentiment Analysis · Natural Language Processing · Language Model · Transfer Learning.

1 Introduction

Sentiment analysis allows us to perform an automated analysis of millions of reviews. With the rapid growth of Twitter, Facebook, Instagram and online review sites, sentiment analysis draws growing attention from both research and industry communities [16]. While it has been extensively researched since 2002 [13], it is still one of the most active research areas in Natural Language Processing (NLP), web mining and social media [25].

Polarity detection is the basic task in sentiment analysis[26]. This task allows us to determine whether a given opinion is positive, negative or neutral. Nowadays, this text classification problem is usually addressed by machine learning methods. Thus, training data and labelled reviews are used to define a classifier [13]. This machine learning approach relies heavily on feature engineering, although recent deep learning algorithms perform this task automatically [12].

* This work was funded by CONCYTEC-FONDECYT under the call E041-01 [contract number 34-2018-FONDECYT-BM-IADT-SE].

In this paper we tackle the polarity detection task using a combined approach of Transfer Learning and Deep Learning. Moreover, we apply this approach on automated classification of Twitter messages in Spanish and its variant Peruvian Spanish (Spanish-PE). This is challenging because of the limited contextual information that Tweets normally contain.

Nowadays, in NLP it is common to use text input encoded as word embeddings such as Word2vec [18], Glove [22] and FastText [1]. When we reuse pre-trained word embeddings in several tasks, we are indirectly employing a transfer learning scheme. In our work, we focus on another transfer learning approach: by pre-training a complete language model for a given language and then use it in text classification, we expect to transfer "knowledge" about the language that allows us to improve the task at hand. In this context, the Universal Language Fine-Tuning (ULMFiT) algorithm has proven to have an impressive performance on several English text classification tasks [10].

Our approach is based on ULMFiT within a deep learning architecture. This setup, which is novel for Spanish sentiment analysis, can be useful in several domains. Overall, our goal is to provide a general procedure that can be applied with less effort in other text classification works. The Deep Learning architecture used is composed by a Recurrent Neural Network [9] and a final dense layer. Those design choices allow us to obtain state-of-the-art results, in terms of F1, over the InterTASS 2017 dataset and competitive results over InterTASS 2018. These datasets were proposed in the TASS workshop at SEPLN. In the last seven years, this workshop has been the main source for Spanish sentiment analysis datasets and proposals [7, 15, 14].

The rest of the paper is organized as follows. In Section 2, related works are explained. Our methodology is presented in Section 3. Experiments and Results are described in Section 4. Finally, Section 5 concludes the paper and Section 6 shows some directions to future work.

2 Related Work

There is a plethora of related works regarding sentiment analysis. However, in this section we are only concerned with contributions for the Spanish language. Arguably one of the most complete Spanish sentiment analysis systems was proposed by Brooke et al. [2], which had a linguistically approach. Recent successful approaches for Spanish polarity classification have been mostly based on machine learning [6].

In the last seven years, the TASS at SEPLN Workshop has been the main source for Spanish sentiment analysis datasets and proposals [7, 15]. Benchmarks for both the polarity detection and aspect-based sentiment analysis tasks have been proposed in several editions of this Workshop (Spanish Tweets have been emphasized).

Recently, deep learning approaches emerge as powerful computational models that discover intricate semantic representations of texts automatically from data without feature engineering. These approaches have improved the state-

of-the-art in many sentiment analysis tasks including sentiment classification of sentences/documents, sentiment extraction and sentiment lexicon learning [25]. However, these results have been mostly obtained for the English Language.

Since 2015, there has been several Deep Learning architectures used for Spanish Twitter Sentiment Analysis, ranging from Multilayer Perceptron (MLP) [11], Recurrent Neural Networks (RNN) [8] and Convolutional Neural Networks (CNN) [24], to name a few. We refer to [19] in order to get an in-depth review of several deep Learning approaches for the Spanish language.

Our proposal is a deep learning approach but, unlike previous approaches, it uses a pre-trained language model to improve the polarity detection task. This setup is novel for the Spanish language.

3 Methodology

The process of Sentiment Analysis using Transfer Learning either of the type of Unsupervised, Inductive or Transductive as described by [21], comprises two steps:

1. The training of a first model on a source domain.
2. The reuse or adaptation of the first model within a second model for training a target domain.

In this sense, the classification task is divided in two sub-tasks, each one with its corresponding model and objective domain (also called dataset). Both of them are related by the weights that the first one provides to the second one when the transference occurs.

In this paper, the Inductive Transfer Learning approach will be used due to the source domain doesn't have labels whereas the target domain does. The source domain has a big dataset that can be used for training a model whose parameters will be after used for improving the training over the target domain. Moreover, in Sentiment Analysis, the source domain should encompass an exhaustive dataset and the target domain usually comprises a specific, small labelled dataset that allow us to classify sentences.

Recent works on text classification [10] highlight the use of an intermediate step which re-trains the first model using sentences of the target domain or a similar one. This process is called Fine-Tuning.

Thus, our pipeline is defined as follows:

– A first model corresponding to a Language Model (LM) which will be trained using sentences from the source domain so as to predict new words. The aim is to learn language essence and to extract deep information about sentence's composition.
  Also, due to the sequential nature of this sub-task [23], a good choice is resort to a Recurrent Neural Network (RNN) architecture. In particular, a suitable model is the weight-dropped AWD-LSTM [17]. AWD-LSTM is an architecture of stacked multi-layer LSTM which uses DropConnect on hidden-to-hidden weights so as to perform a recurrent regularization. In addition, a

variation of the averaged stochastic gradient method, called NT-ASGD, is used.
– An intermediate step for fine-tuning which re-trains the parameters of the first model using the target domain.
– A second model which will be adapted from the first one, adding two layers for classification. This model takes advantage of the knowledge learned and will be trained on the labelled target domain.

This methodology to address general text classification problems [10] also proposes techniques such as gradual unfreezing, discriminative fine-tuning (Discr) and slanted triangular learning rate (STLR) for dealing with overfitting. The overall process is called: Universal Language Fine-Tuning (ULMFiT).

The initial pipeline is updated using ULMFiT, as shown in Figure 1. The final three stages are:

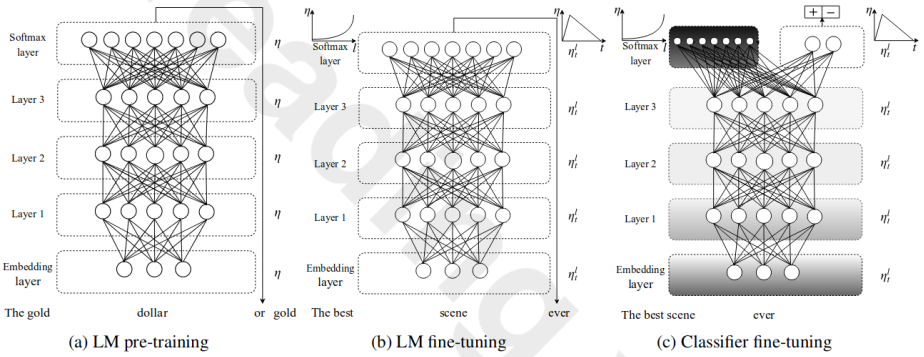


Fig. 1: Stages of updated pipeline using ULMFiT [10].

(a) The language model (LM) is trained on a general domain corpus to capture general features of the language in different layers.
(b) The full LM is fine-tuned on target task data using discriminative fine-tuning (Discr) and slanted triangular learning rates (STLR) to learn task-specific features.
(c) The classifier is fine-tuned on the target task using gradual unfreezing, Discr, and STLR to preserve low-level representations and adapt high-level ones (shaded: unfreezing stages; black: frozen).

This pipeline has shown to improve the state-of-the-art text classification on several datasets for English language. In this work, we have adopted the ULMFit pipeline for Sentiment Analysis on Spanish tweets datasets. Experiments show effectiveness of this approach for this task.

4 Experiments

A detailed setup about the hardware and software requirements for reproducing this paper are described in this section. In addition, we show hyperparameters tuned during experimentation by picking a learning rate that lead to convergence without overfitting and regularization.

4.1 Technical Resources

All experiments were carried out in Jupyter Notebooks running Python 3.6 kernel and Pytorch 0.3.1.

For a complete detail about dependencies used, the repository of the project is available at [20].

All models were trained on a Google Cloud VM with 2 vcpu, 13 GB of RAM and GPU K80 with 12 GB GDDR5.

4.2 Dataset

To train the entire model end to end, three data sources were used:

- **The General Language Model** was trained on a dump of the entire Spanish Wikipedia, from which only the top 100 million articles were kept and the vocabulary was limited to 60,000 tokens in accordance to the English setting approach.
- **The Specific Language Model** was trained on 136,286 unlabeled tweets for the Spanish language which were collected from Twitter using Tweepy (3.7.0).
- **The Specific Task Classifier** was trained on the dataset published by The Spanish Society for Natural Language Processing (SEPLN) for InterTASS (Task1) Competition 2017 [15] and InterTASS-PE (Task1 / Sub-task 2) Competition 2018 [14].
  The summarized data is shown in Tables 1a and 1b.

Table 1: Tweets distribution over InterTASS datasets.

(a) InterTASS 2017.

	Training	Development	Test
P	317	156	642
NEU	133	69	216
N	416	219	767
NONE	138	62	274
Total	1008	506	1899

(b) InterTASS-PE 2018.

	Training	Development	Test
P	231	95	430
NEU	166	61	367
N	242	106	472
NONE	316	238	159
Total	1000	500	1428

4.3 Pre-Processing

Each corpus was pre-processed as follows:

- Twitter user references were replaced by the token "user_ref".
- URL references were replaced by the token "hyp_link".
- Hashtags comments were replaced by the token "hash_tag".
- Slang words were replaced by their formal equivalent, for example, "q" and "k" were replaced by the correct word "que".
- Interjections denoting laughter ("jaja", "jeje", "jiji", "jojo") were replaced by the token "risa_ja", "risa_je", "risa_ji", "risa_jo".
- Any other characters like "\n", "&lt", "&gt", "\xa0" were replaced by a space character.
- Redundant space character were removed.
- The text was converted to lowercase.

4.4 Architecture

The architecture for the General Language Model as well as the Specific Language Model [10] were comprised by a three-layer LSTM model with 1150 units in the hidden layer and embedding size of 400.

In the classifier, two linear blocks with batch normalization and dropout were added to the previous model. Rectified Linear Unit activations were used for the intermediate layer. A Softmax activation was used in the last layer.

4.5 Hyperparameters

For both training datasets, InterTASS (Task1) Competition 2017 [15] and InterTASS-PE (Task1 / Sub-task 2) Competition 2018 [14], the hyperparameters are similar across all stages of the ULMFiT method.

The main hyperparameters shared for all models were:

- Backpropagation Through Time (BPTT): 70
- Weight Decay (WD): $1e-7$
- The batch size (BS) was limited by the available GPU memory.

Besides these parameters, the models used different configurations for the learning rate (LR), dropouts, cyclical learning rate (CLR) and slanted triangular learning rates (STLR). Additionally, gradient clipping (CLIP) was applied to the models.

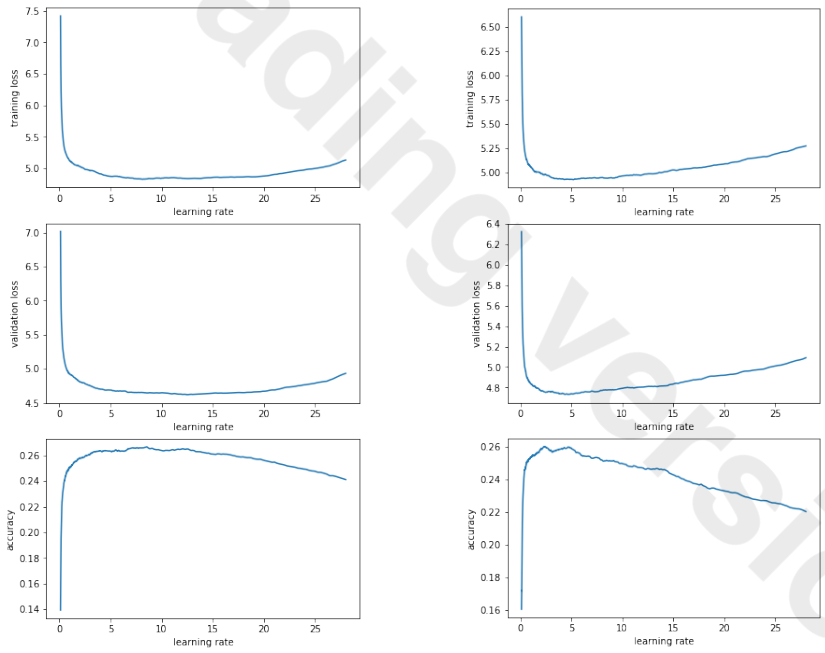
Two configurations of dropout were used (see Table 2).

Table 2: Dropout configurations.

Dropout	ULMFiT [10]	AWD-LSTM [17]
Embedding Dropout	0.02	0.10
Input Dropout	0.25	0.60
Weight Dropout	0.20	0.50
Hidden Dropout	0.15	0.20
Output Dropout	0.10	0.40

General Language Model The hyperparameters for this model were directly transferred according Howard and Ruder work [10]. Scripts used were taken from the official Fastai repository [4] and ULMFiT repository [5].

Specific Language Model A Learning Rate Finder (LRF) was used to determine suitable candidate learning rates, several values are depicted in Figure 2.



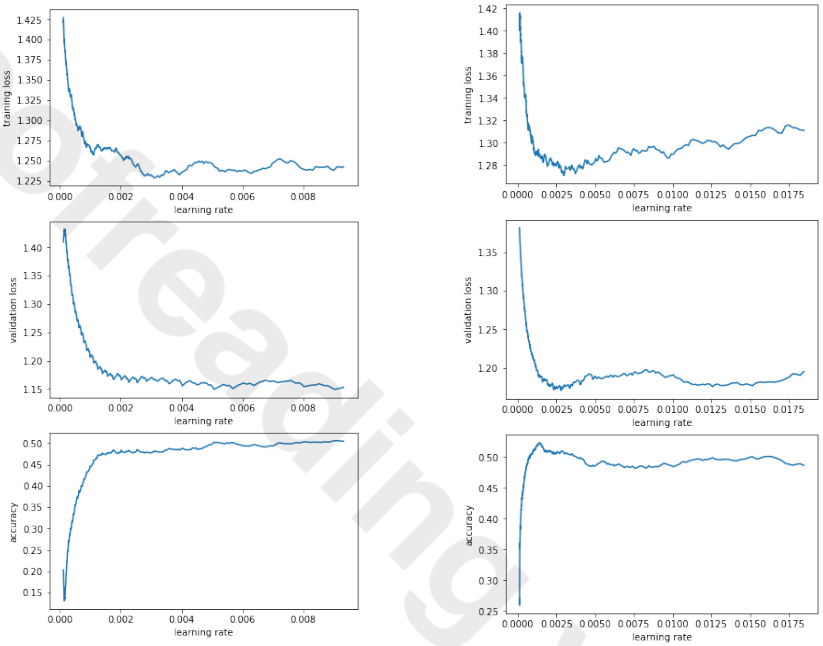
(a) InterTASS 2017. (b) InterTASS-PE 2018.

Fig. 2: LRF for Specific Language Model.

With the information extracted from Figure 2, a suitable learning rate (LR) was set to 12 for InterTASS 2017 and 5 for InterTASS-PE 2018 dataset.

Also, the gradient clipping (CLIP) was set to 0.92 and the dropout configuration (Table 2) was set to 0.8∗ULMFiT.

Specific Task Classifier Similar to the previous model, a learning rate finder (LRF) was used to determine suitable candidate learning rate (Figure 3).



(a) InterTASS 2017.

(b) InterTASS-PE 2018.

Fig. 3: LRF for Specific Task Classifier

With the information extracted from Figure 3, a suitable LR was set to $3e-3$ for InterTASS 2017 and $2.5e-3$ for InterTASS-PE 2018 dataset.

Also, the gradient clipping (CLIP) was set to 0.12 and the dropout configuration (Table 2) was set to 1.0∗ULMFiT.

4.6 Results

The results for InterTASS (Task1) Competition 2017 [15] were better than expected as shown in Table 3a, achieving the second best result, according to M-F1 metric (the ELiRF-UPV team reached a M-F1 score of 0.493).

Likewise, results on InterTASS-PE (Task1 / Sub-task 2) Competition 2018 [14], are shown in Table 3b. While they weren't the best, they are within the best nine results of the competition.

Table 3: Results over InterTASS Test datasets.

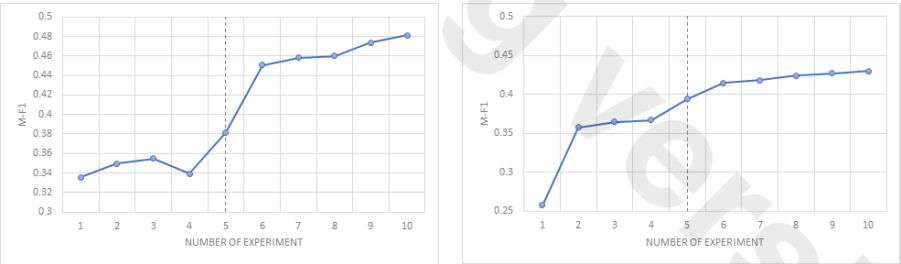
(a) InterTASS 2017.

Team	M-F1	Acc.
ELiRF-UPV-run1	0.493	0.607
Our proposal	**0.481**	**0.567**
RETUYT-svm_cnn	0.471	0.596
ELiRF-UPV-run3	0.466	0.597
ITAINNOVA-model4	0.461	0.476
jacerong-run-2	0.460	0.602
jacerong-run-1	0.459	0.608
INGEOTEC-evodag001	0.457	0.507
RETUYT-svm	0.457	0.583
tecnolengua-sentonly	0.456	0.582

(b) InterTASS-PE 2018.

Team	M-F1	Acc.
retuyt-cnn-pe-1	0.472	0.494
atalaya-pe-lr-50-2	0.462	0.451
retuyt-lstm-pe-2	0.443	0.488
retuyt-svm-pe-2	0.441	0.471
ingeotec-run1	0.439	0.447
elirf-intertass-pe-run-2	0.438	0.461
atalaya-mlp-sentiment	0.437	0.520
retuyt-svm-pe-1	0.437	0.474
Our proposal	**0.436**	**0.463**
elirf-intertass-pe-run-1	0.435	0.440

As depicted in Figures 4a and 4b, the first 5 experiments on both benchmarks, during the fine-tuning process, used only train and development datasets for training the Specific Language Model which turned out poor results. Conversely, the next 5 experiments included tweets extracted using the Twitter API, as described in subsection 4.2. It is worth noting the improvement obtained after this addition.



(a) InterTASS 2017. (b) InterTASS-PE 2018.

Fig. 4: Evolution of results in Test dataset for InterTASS Competitions. The gray line indicates a before and after including the tweets extracted using the Twitter API for training the Specific Language Model.

In case of InterTASS-PE 2018, although there is a clear improvement, it is less impressive than InterTASS 2017. Perhaps, a tailored selection of Peruvian Tweets for training its Specific Language Model would have returned better results.

5 Conclusions

We have adapted ULMFit in order to perform sentiment analysis on Spanish Tweets — This approach is novel for this language. By using this setup we have achieved competitive results considering that the SEPLN datasets [15, 14] are very challenging due to the limited contextual information provided.

Our best result (comparable to the state-of-the-art) was obtained on the InterTASS 2017 dataset. In contrast, we have experienced some difficulties using the InterTASS-PE 2018 dataset which is composed by Tweets in Spanish from the Peruvian dialect. These difficulties are due to the text used for training the Specific Language Model was mostly extracted from Castilian Spanish tweets. Thus, it is worth noting that training or fine-tuning the Specific Language Model with a relevant dataset has a positive impact in the classification task at hand.

6 Future Work

Recently, another transfer learning approach, called Bidirectional Embedding Representations from Transformers (BERT) [3], has emerged as the state-of-the-art in many NLP tasks. The use of this language model could enhance the performance of the pipeline presented in this work.

Also, there is a room for improvement if the language model approach is combined with data augmentation techniques. In this sense, new approaches for data augmentation such as the "BERT contextual augmentation" algorithm [27] could help us to improve our results if applied to our smalls datasets.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
2. Brooke, J., Tofiloski, M., Taboada, M.: Cross-linguistic sentiment analysis: From english to spanish. In: Proceedings of RANLP 2009. pp. 50–54 (2009)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv e-prints **abs/1810.04805**, arXiv:1810.04805 (Oct 2018)
4. Fastai: Fastai (May 2019), https://github.com/fastai/fastai
5. Fastai: Ulmfit (May 2019), https://github.com/fastai/fastai/tree/master/courses/dl2/imdb_scripts
6. Garcia, M., Martinez, E., Villena, J., Garcia, J.: Tass 2015 – the evolution of the spanish opinion mining systems. Procesamiento de Lenguaje Natural **56**, 33–40 (2016)
7. Garcia-Cumbreras, M.A., Villena-Roman, J., Martinez-Camara, E., Diaz-Galiano, M., Martin-Valdivia, T., Ureña Lopez, A.: Overview of tass 2016. In: Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN. pp. 13–21 (2016)
8. Garcia-Vega, M., Montejo-Raez, A., Diaz-Galiano, M.C., Jimenez-Zafra, S.M.: Sinai in tass 2017: Tweet polarity classification integrating user information. In: Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN. pp. 91–96 (2017)

9. Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, Springer, Berlin (2012), https://cds.cern.ch/record/1503877
10. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://www.aclweb.org/anthology/P18-1031
11. Hurtado, L.F., Pla, F., Gonzalez, J.A.: Elirf-upv at tass 2017: Sentiment analysis in twitter based on deep learning. In: Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN. pp. 29–34 (2017)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521(7553)**, 436–444 (2015)
13. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan and Claypool Publishers (2012)
14. Martinez-Camara, E., Almeida-Cruz, Y., Diaz-Galiano, M., Estévez-Velarde, S., Garcia-Cumbreras, M.A., Garcia Vega, M., Gutiérrez, Y., Montejo-Ráez, A., Montoyo, A., Munoz, R., Piad-Morffis, A., Villena-Roman, J.: Overview of tass 2018: Opinions, health and emotions. In: Proceedings of TASS 2018: Workshop on Sentiment Analysis at SEPLN. pp. 13–27 (2018)
15. Martinez-Camara, E., Diaz-Galiano, M., Garcia-Cumbreras, M.A., Garcia-Vega, M., Villena-Roman, J.: Overview of tass 2017. In: Proceedings of TASS 2017: Workshop on Sentiment Analysis at SEPLN. pp. 13–21 (2017)
16. McGlohon, M., Glance, N., Reiter, Z.: Star quality: Aggregating reviews to rank products and merchants. In: Proceedings of Fourth International Conference on Weblogs and Social Media (ICWSM) (2010)
17. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing LSTM language models. CoRR **abs/1708.02182** (2017), http://arxiv.org/abs/1708.02182
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013), http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
19. Ochoa-Luna, J., Ari, D.: Deep neural network approaches for spanish sentiment analysis of short texts. In: Simari, G.R., Fermé, E., Gutiérrez Segura, F., Rodríguez Melquiades, J.A. (eds.) Advances in Artificial Intelligence - IBERAMIA 2018. pp. 430–441. Springer International Publishing, Cham (2018)
20. Palomino, D.: Ulmfit implementation for tass dataset evaluation (May 2019), https://github.com/dpalominop/ULMFit
21. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (Oct 2010). https://doi.org/10.1109/TKDE.2009.191
22. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
23. Rother, K., Rettberg, A.: Ulmfit at germeval-2018: A deep neural language model for the classification of hate speech in german tweets. Proceedings of the GermEval 2018 Workshop pp. 113–119 (2018)

24. Segura-Bedmar, I., Quiros, A., Martínez, P.: Exploring convolutional neural networks for sentiment analysis of spanish tweets. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1014–1022. Association for Computational Linguistics (2017), http://aclweb.org/anthology/E17-1095
25. Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., Zhou, M.: Sentiment embeddings with applications to sentiment analysis. Knowledge and Data Engineering, IEEE Transactions on **28**(2), 496–509 (Feb 2016)
26. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1073083.1073153, http://dx.doi.org/10.3115/1073083.1073153
27. Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional BERT contextual augmentation. CoRR **abs/1812.06705** (2018), http://arxiv.org/abs/1812.06705